Statistical Induction, Severe Testing, and Model Validation

Aris Spanos*
Virginia Tech, Department of Economics,
Blacksburg, VA 24061, USA

June, 2006

Abstract

A number of important methodological issues in statistical modeling and inference depend crucially on the notion of statistical induction adopted. An attempt is made to articulate the notion of statistical induction underlying modern frequentist inference going back to Fisher (1922). The paper brings out the differences in the nature of inductive inference associated with estimation and prediction on one hand and testing on the other; the former based on factual and and the latter on counterfactual reasoning. Induction by enumeration is placed in a formal statistical context in order to bring out its crucial weaknesses.

Particular emphasis is placed on the role of pre-data type I and II error probabilities, as measures of the 'trustworthiness' of test procedures. Post-data, error probabilities can be used to render the traditional coarse accept/reject decision more informative by evaluating the severity with which a hypothesis or a claim passes a particular test, with data $\mathbf{x}$. The discussion emphasizes the nature of the severity assessment and the associated post-data error probabilities, as they relate to the pre-data error probabilities. Supplementing N-P testing with the severity evaluation gives rise to the error-statistical account of inference which constitutes the most complete description of frequentist statistical induction.

The evaluation of error probabilities (pre-data and post-data) assumes the validity of the statistical premises, because any departures will render the inductive inference unreliable by creating a divergence between the nominal and actual error probabilies. The paper discusses the importance of ensuring statistical adequacy using thorough misspecification testing and respecification. It also demonstrates how statistical adequacy can be used to shed light on a number of methodological problems such as model validation vs. model selection and statistical vs. substantive adequacy.

*Section 3 of the paper relies heavily on joint work with Deborah Mayo.

1 Introduction

A number of important methodological issues in statistical inference and modeling, including criticisms of Neyman-Pearson testing and Bayesian vs. frequentist methodological debates, depend crucially on the notion of statistical induction adopted. In view of the fact that the conception of statistical induction is often implicit in statistical discussions, an attempt is made to articulate the notion underlying the frequentist approach to inference going back to Fisher (1922), and consider a number of methodological issues that revolve around statistical induction. Particular emphasis is placed on:

(i) the form and structure of the premises of induction,

(ii) the nature of 'objectivity' in frequentist inference,

(iii) the role of pre-data error probabilities in determining the optimality of an inference procedure,

(iv) the role of post-data error probabilities in providing an inferential interpretation of the accept/reject decision, and

(iv) the role of statistical adequacy in ensuring the reliability of the inference.

The aim of statistical modeling and inference is to learn about certain aspects of real world phenomena exhibiting chance regularity. Elucidating the form and structure of the premises of statistical induction is particularly important for a variety of reasons including the fact that it holds the key to resolving certain chronic methodological problems in statistics. This can also shed light on the claim that defining these premises in terms of probabilistic assumptions invariably involves subjective judgements which taint the objectivity of inference. It is argued that the statistical premises can and should be specified in terms of the probabilistic structure of the observable processes involved, rendering the assumptions testable vis-a-vis a particular data $\mathbf{x} =(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

It is long-familiar that pre-data the type I and II error probabilities provide a way to appraise the probativeness of the test procedure – its capacity of to detect discrepancies from the null hypothesis – and define its optimality. What is less well-known is that error probabilities can be used post-data to render the coarse Neyman-Pearson (N-P) accept/reject decision more informative by evaluating the severity with which a hypothesis or a claim passes a particular test, with given data $\mathbf{x}$. The same post-data error probabilities can be used to guard against perpetrating the well-known fallacies of acceptance and rejection.

A crucial precondition for the cogency of error probabilities (pre and post-data) is the validity of the statistical premises vis-a-vis data $\mathbf{x}$. Any departure from the premises will render the inductive inference unreliable to a greater or lesser extent, by inducing a difference between the *nominal* and *actual* error probabilities. The paper reflects on this problem and discusses different ways to address the statistical adequacy issue. The recommendation is thorough misspecification testing and discerning respecification to account for the systematic information in the data. Using statistical

adequacy the paper brings out certain weaknesses in model selection procedures and discusses the difference between statistical and substantive adequacy.

2 Statistical Induction: pre-data

2.1 Early 20th century

For Karl Pearson statistical modeling would begin with data $\mathbf{x} := (x_1, x_2, ..., x_n)$ in search of a descriptive model which would be in the form of a frequency curve $f(x; \widehat{\boldsymbol{\theta}})$, chosen from the Pearson family $f(x; \boldsymbol{\theta})$, after applying the method of moments to estimate $\boldsymbol{\theta}$ (see Pearson, 1895).

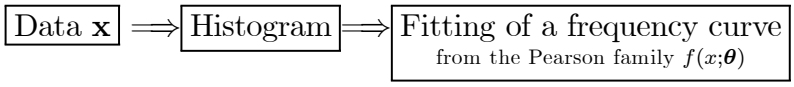

Fig. 1: The Pearson method of moments

The problem of *statistical induction* was understood by Karl Pearson (1920) in terms of being able to ensure the 'stability' of the estimation results in subsequent samples, by invoking 'uniformity' and 'representativeness' assumptions. This is a form of *induction by enumeration,* which attempts to generalize observed *events:*

'80% of A's in this data are B's', beyond the data in hand; see Salmon (1967).

2.2 Fisher's initiating the recasting of statistical induction

Fisher's most enduring contribution to modern statistics is his pioneering the recasting of *statistical induction.* Instead of starting with data $\mathbf{x}$ in search of a descriptive model, he would interpret the data as *a representative sample* from a pre-specified 'hypothetical infinite population'; Fisher (1922, 1925). This might seem like a trivial re-arrangement of Karl Pearson's procedure, but in fact constitutes a complete reformulation of statistical induction from generalizing observed 'events' relating to the data, to modeling the underlying 'process' that gave rise to the data. This way Fisher was able to quantify the uncertainty associated with the inductive inference in the form of 'ascertainable error probabilities'.

In particular, Fisher devised a general way to quantify this uncertainty by (a) *embedding* the *material experiment* into a *statistical model*, and then (b) use the latter to ascertain the *(frequentist) error probabilities* associated with particular inferences in its context. These error probabilities are *deductively* derived from the statistical model, and provide a measure of the 'trustworthiness' of the inference procedure: how often a certain method will give rise to reliable inferences concerning the underlying actual Data Generating Mechanism (DGM). The form of induction envisaged by Fisher, and even earlier by Peirce (see Mayo, 1996, ch. 12), is one where the reliability

of the inference stems from the 'trustworthiness' of the procedure used to arrive at the inference. As argued by (Fisher, 1935, p. 14):

"In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure."

The inference is reached by an inductive procedure which, with high probability, will reach true conclusions (estimation, testing, prediction) from true (or approximately true) premises (statistical model).

For Fisher the process of statistical modeling begins with a *prespecified parametric statistical model* $\mathcal{M}$ ('a hypothetical infinite population'), chosen to ensure that the observed data $\mathbf{x}$ can be viewed as a *random sample* from that 'population':

"The postulate of randomness thus resolves itself into the question, "Of what population is this a random sample?" which must frequently be asked by every practical statistician." (Fisher, 1922, p. 313)

Fisher initiated the recasting of statistical induction that rendered its *premises testable*, by viewing data $\mathbf{x}$ as a 'random sample' from a prespecified population. Using Neyman's generalization of Fisher's notion of a statistical model from an 'infinite population' to a 'chance mechanism' defined in terms of a generic stochastic process $\{X_k,\ k\in\mathbb{N}\}$, and extending the notion of data $\mathbf{x}$ from being a 'random sample' to being a 'truly representative' realization of such a process, one can define the *Fisher-Neyman probabilistic perspective* (see Spanos, 2006a-c). From this perspective a statistical model is viewed as a *parameterization* of a generic stochastic process $\{X_k,\ k\in\mathbb{N}\}$ whose probabilistic structure is such that would render $\mathbf{x}$ 'a truly typical realization' thereof. To Fisher we also owe the notion of a parameter and a parametric statistical model.

Fisher also recognized the fact that the *trustworthiness* of an inference procedure depends crucially on the *adequacy* of the assumed statistical model vis-a-vis data $\mathbf{x}$, and suggested that it be tested:

"As regards problems of specification, ... the adequacy of our choice may be tested *a posteriori*." (p. 314)

Statistical adequacy secures the reliability of inference by ensuring that the actual error probabilities approximate closely the nominal error probabilities. It is interesting to note that the first tests discussed by Fisher (1925) were *misspecification tests* concerned with assessing the Normality, Independence and Identically Distributed assumptions.

In summary, Fisher pioneered the recasting of statistical induction in terms of 'reliable procedures' based on 'ascertainable error probabilities'. The Fisher-Neyman perspective provides a purely probabilistic construal of statistical models which can be used to disentangle the respective roles of *substantive* and *statistical information* in empirical modeling (see Lehmann, 1990, Cox, 1990). Although this issue is beyond the scope of the present paper (see Spanos, 2006b), in what follows it is important to note that in the context of the Probabilistic Reduction approach the roles of the statistical and substantive information are considered complementary. Hence, *ab initio,*

the statistical information is captured by the statistical model and the substantive information by a *structural model*. The connection between the two models is that a structural model acquires statistical operational meaning when embedded into an adequate statistical model. This perspective can been used to shed light on a number of methodological issues relating to specification, misspecification testing, and respecification, including the role of preliminary data analysis, structural vs. statistical models, model specification vs. model selection, and statistical vs. substantive adequacy; see Spanos (2006a-c).

2.3 The notion of a statistical model

The cornerstone of modern statistics is the notion of a *statistical model* which can be viewed as an internally consistent set of probabilistic assumptions aiming to provide an 'idealized' (statistical) description of the stochastic mechanism that gave rise to the observed data $\mathbf{x} := (x_1, x_2, ..., x_n)$. The quintessential example is the *simple Normal model* given in table 1, comprising a statistical Generating Mechanism (GM), and the probabilistic assumptions [1]-[4].

Table 1 - Simple Normal Model

Statistical GM: $X_k = \mu + u_k,\ k \in \mathbb{N},$

$$\begin{array}{lll} [1] & \text{Normality:} & X_k \backsim \mathsf{N}(.,.), \\ [2] & \text{Mean homogeneity:} & E(X_k) := \mu, \\ [3] & \text{Variance homogeneity:} & Var(X_k) := \sigma^2, \end{array} \Big\} \ k \in \mathbb{N}. \qquad (1)$$

[4] Independence: $\{X_k,\ k \in \mathbb{N}\}$ k-independent process

Viewed from the Fisher-Neyman probabilistic perspective this model represents a parameterization of the generic process stochastic process $\{X_k,\ k\in\mathbb{N}\}$ assumed to be *Normal, Independent and Identically Distributed* (NIID). This suggests that this is an appropriate statistical model to select in cases where data $\mathbf{x} =(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ can be realistically viewed as a realization of a generic NIID process.

To get some idea of what this selection involves consider figures 1-4 where figure 1 represents a typical realization of a NIID process, but figures 2-4 do not. In particular, figure 2 shows a realization of a non-Normal (skewed) IID process, figure 3 exhibits a realization of a Normal, Independent but non-ID (trending mean) process, and figure 4 depicts a typical realization of a Normal, Markov and Stationary process; the cycles indicate positive Markov dependence; see Spanos (1999), ch. 5.

More formally the relationship between the stochastic process $\{X_t,\ t \in \mathbb{T}\}$ and the simple Normal model in table 1, can be expressed in the form of a reduction that connects the *joint distribution* $D(X_1, X_2, ..., X_n; \boldsymbol{\varphi})$ to the distribution underlying the model $D(X_k; \boldsymbol{\phi}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(X_k - \mu)^2\}$ via the probabilistic assumptions of

NIID:

$$D(X_1, X_2, ..., X_n; \boldsymbol{\varphi}) \overset{\mathsf{I}}{=} \prod_{k=1}^{n} D_k(X_k; \boldsymbol{\phi}_k) \overset{\mathsf{IID}}{=} \prod_{k=1}^{n} D(X_k; \boldsymbol{\phi}). \tag{2}$$

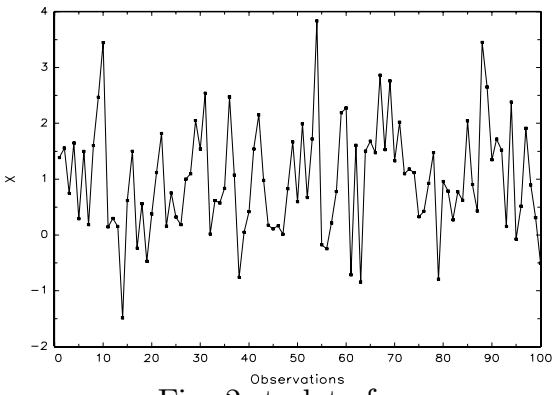

Fig. 2: t-plot of x_t

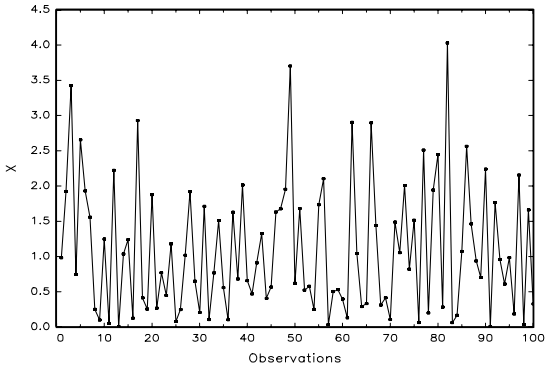

Fig. 3: t-plot of x_t

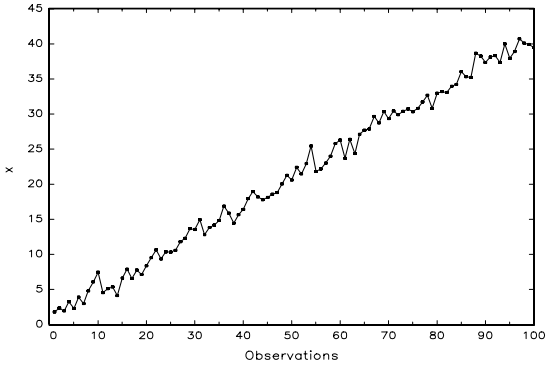

Fig. 4: t-plot of x_t

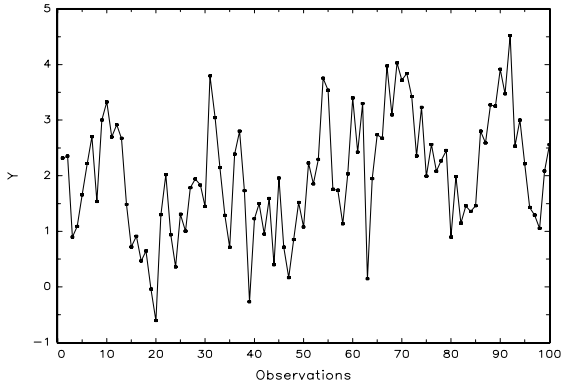

Fig. 5: t-plot of x_t

The form and structure of a statistical model is specified entirely in terms of probabilistic concepts that relate directly to the *joint distribution* $D(X_1, X_2, ..., X_n; \boldsymbol{\varphi})$. The reduction (probabilistic) assumptions come from three broad categories, *Distribution, Dependence and Heterogeneity,* and specification can be viewed as the result of partitioning the set of all possible models that could have given rise to the data; see figure 6.

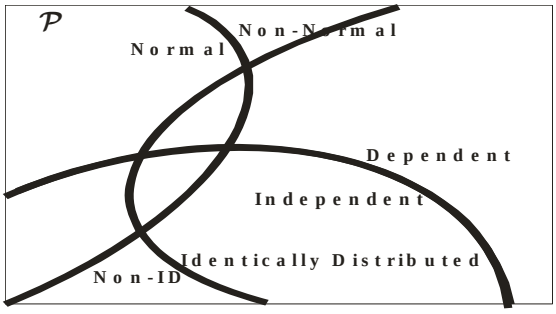

Fig. 6 - Specification by partitioning

The statistical model is often denoted by:

$$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) = \{f(\mathbf{x};\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\},\ \mathbf{x} \in \mathcal{X} := \mathbb{R}_X^n,\ n > 1,$$

and viewed as a subset of $\mathcal{P}(\mathbf{x})$ the set of all possible statistical models that could have given rise to data $\mathbf{x}_0$, i.e. $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}) \subset \mathcal{P}(\mathbf{x})$. $\mathcal{P}(\mathbf{x})$ provides the wider context for statistical modeling because it brings out the fact that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is one of many (possibly infinite) statistical models which is characterized by the probabilistic structure attributed to the underlying process $\{X_k,\ k \in \mathbb{N}\}$.

By imposing different probabilistic assumptions from three broad categories, *Distribution, Dependence and Heterogeneity*, on the process $\{X_k,\ k \in \mathbb{N}\}$, one can derive numerous statistical models belonging to the set $\mathcal{P}(\mathbf{x})$, via reductions analogous to (25); see Spanos (1986, 1995, 1999). In the context of $\mathcal{P}(\mathbf{x})$ one can view the problem of *specification* as the choice of the probabilistic assumptions for $\{X_k,\ k \in \mathbb{N}\}$ that would render data $\mathbf{x}_0$ a truly typical realization thereof. The problem of *Mis-Specification* (M-S) *testing* can be viewed as probing the complement $\mathcal{P}(\mathbf{x}) - \mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ for possible departures, and that of *respecification* as choosing a more appropriate statistical model within $\mathcal{P}(\mathbf{x})$, when $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is found wanting.

2.4 Statistical inference

Statistical inference constitutes a special form of inductive inference with its particularity emanating from the form of its premises. The *premises* of statistical induction comprise a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \mathbf{x} \in \mathcal{X}$ and a particular data $\mathbf{x}_0$, where $\mathbf{x}_0$ is viewed as a 'truly typical' realization of the stochastic process $\{X_k,\ k \in \mathbb{N}\}$ underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. The *inference* is framed in terms of the unknown parameter(s) $\boldsymbol{\theta} \in \Theta$, but it is, in the final analysis, concerned with the underlying stochastic mechanism that gave rise to data $\mathbf{x}_0$. The choice of the statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \boldsymbol{\theta} \in \Theta$ amounts to reducing the set of all possible statistical models $\mathcal{P}(\mathbf{x})$ to a small subset, and inductive inference is concerned with narrowing that subset even further in an attempt to determine the 'true' stochastic mechanism $\mathcal{M}_{\boldsymbol{\theta}_*}(\mathbf{x})$, which is a point in $\mathcal{P}(\mathbf{x})$; $\boldsymbol{\theta}_*$ being the true value of $\boldsymbol{\theta}$.

Statistical inference in frequentist statistics takes different forms:

(i) point estimation, (ii) interval estimation, (iii) hypothesis testing, and (iv) prediction.

All forms of frequentist inference suppose at the outset that the 'true' DGM belongs to the prespecified family of models $\{\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \boldsymbol{\theta} \in \Theta,\ \mathbf{x} \in \mathcal{X}\}$. Hence, it should come as no surprise to learn that these inference procedures are based on mappings of the form:

$$g(.,.) :\ [\Theta \times \mathcal{X}] \to \mathbb{R},$$

in conjunction with the probabilistic structure of the model as summarized by the *distribution of the sample:*

$$f(\mathbf{x};\boldsymbol{\theta}),\ \mathbf{x} := (x_1, ..., x_n) \in \mathcal{X}.$$

A point estimator of θ, say $\widehat{\theta} = h(X_1, X_2, ..., X_n)$, can viewed as a mapping:

$$h(.) : \mathcal{X} \to \Theta,$$

with a sampling distribution:

$$F(\theta) = \mathbb{P}(h(X_1, ..., X_n) \leq \theta) = \overbrace{\int\int \cdots \int}^{\text{n times}}_{\{\mathbf{x}:\, h(\mathbf{x}) \leq \theta\}} f(x_1, ..., x_n; \theta)\, dx_1 \cdots dx_n, \text{ for all } \theta \in \Theta. \tag{3}$$

The idea is to use the data $\mathbf{x}_0$ to select the most representative value $\widehat{\theta} = h(\mathbf{x}_0)$ for θ. That yields the estimated model $\mathcal{M}_{\widehat{\boldsymbol{\theta}}}(\mathbf{x}),\ \mathbf{x} \in \mathcal{X}$ as the one element of the prespecified family of models selected by data $\mathbf{x}_0$. Similarly, a test statistic associated with a null hypothesis $\boldsymbol{\theta} \in \Theta_0 \subset \Theta$, say $T = h(X_1, X_2, ..., X_n; \boldsymbol{\theta}_0)$, can viewed as a distance mapping:

$$h(.) : \mathcal{X} \to \mathbb{R},$$

with a sampling distribution $F(t)$ defined analogously to (3). The idea is to pose the question whether data $\mathbf{x}_0$ provide evidence that the true model $\mathcal{M}_{\boldsymbol{\theta}_*}(\mathbf{x})$ belongs to a subset $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, $\boldsymbol{\theta} \in \Theta_0$ of the prespecified model or not. That is, do the data warrant narrowing down the original family of models to the subset $\mathcal{M}_{\boldsymbol{\theta}_0}(\mathbf{x})$?

Example. In the case of the simple Normal model, the statistics:

$$\overline{X} = \tfrac{1}{n}\sum\nolimits_{k=1}^{n} X_k, \qquad s^2 = \tfrac{1}{(n-1)}\sum\nolimits_{k=1}^{n}(X_k - \overline{X})^2, \tag{4}$$

constitute 'good' estimators of (μ, σ^2), with sampling distributions:

$$\overline{X} \backsim \mathsf{N}\left(\mu, \tfrac{\sigma^2}{n}\right), \qquad \tfrac{(n-1)s^2}{\sigma^2} \backsim \chi^2(n-1). \tag{5}$$

Moreover, for testing the *hypotheses*:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0, \tag{6}$$

the test statistic $\tau(\mathbf{X}) = \frac{\sqrt{n}(\overline{X} - \mu_0)}{s}$, with a sampling distribution:

$$\tau(\mathbf{X}) = \tfrac{\sqrt{n}(\overline{X} - \mu_0)}{s} \overset{H_0}{\backsim} \mathsf{St}(n-1),$$

can be used to define an optimal test when combined with the rejection region $C_1(\alpha) = \{\mathbf{x} : \tau(\mathbf{x}) > c_\alpha\}$; see Cox and Hinkley (1974). Using the well-known duality between hypothesis testing and *interval estimation* on can define the two-sided Confidence Interval (CI) for μ :

$$\mathbb{P}\left(\overline{X} - c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}}) \leq \mu \leq \overline{X} + c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}})\right) = 1 - \alpha.$$

Prediction differs from the above inferences in so far as it is concerned with finding the most representative value of X_k beyond the observed data, say X_{n+1}. A good predictor of X_{n+1} is given by:

$$\widehat{X}_{n+1} = h(\overline{X}), \text{ where } h(\overline{X}) = \begin{cases} 1, & \text{if } \overline{X} \geq \frac{1}{2}, \\ 0, & \text{if if } \overline{X} < \frac{1}{2}. \end{cases} \tag{7}$$

2.5 The nature and forms of statistical induction

How do these inference procedures differ? It turns out that their differences arise from the nature of induction involved. The traditional statistical literature distinguishes between estimation (point and interval) and hypothesis testing. Chatterjee (2003) refers to the former as *open induction* and to the latter as *hypothetic induction* paraphrasing approvingly Day (1961): "Some philosophers regard hypothetic induction more important than open induction for the progress in science ..., since one can give free play to one's imagination in framing the hypothesis." (ibid. p. 65)

This is a very interesting point that it is not widely appreciated in statistics. In point estimation one selects the most representative value (in view of the data) of the unknown parameter θ; representativeness being defined in terms of optimal properties such as unbiasedness, efficiency, sufficiency, consistency etc. The main problem with this form of inductive inference is that the error probabilities associated with inferring a point estimate are rather vague, rendering this form of inductive inference less precise and effective. Ensuring that an estimator is consistent, unbiased or even fully efficient does not provide one with enough information to evaluate the reliability of a point estimate inference. Interval estimation remedies this deficiency by providing a way to evaluate the relevant error probabilities associated with an interval estimator covering the true value of the unknown parameter θ. Hypothesis testing poses even more probative questions; whether particular hypothetical values of θ are warranted in view of the data. The questions posed in estimation and testing as well as the answers elicited are different because the form of reasoning underlying these two forms of inferences is dissimilar.

The form of reasoning that underlies estimation is that of *factual reasoning* based on evaluating the sampling distributions of estimators 'under the true state of nature' (TSN):

$$\overline{X} \overset{\text{TSN}}{\backsim} \mathsf{N}\left(\mu_*, \tfrac{\sigma_*^2}{n}\right), \qquad \tfrac{(n-1)s^2}{\sigma_*^2} \overset{\text{TSN}}{\backsim} \chi^2(n-1). \tag{8}$$

That is, the sampling distributions are evaluated assuming that the unknown parameters take their 'true' values, say (μ_*, σ_*^2), whatever those happen to be.

In contrast to estimation, the reasoning underlying hypothesis testing is *counterfactual*. The sampling distribution of a test statistic is evaluated under several hypothetical scenarios based on 'what if' counterfactuals. In particular, what is the sampling distribution of the test statistic if the null or the alternative hypotheses are

true? In the above case these scenarios give rise to:

$$\frac{\sqrt{n}(\overline{X}-\mu_0)}{s} \overset{H_0}{\backsim} \mathsf{St}(n-1), \qquad \frac{\sqrt{n}(\overline{X}-\mu_0)}{s} \overset{H_1}{\backsim} \mathsf{St}(\delta; n-1), \text{ for any } \mu_1 > \mu_0, \tag{9}$$

where $\delta = \frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$. The counterfactual reasoning in testing poses sharper questions by assuming different hypothetical values for μ, and often elicits more informative answers from the data.

This demarcation line between the two types of reasoning is best brought out when one considers the sampling distribution underlying the two-sided confidence interval (CI) for μ :

$$\mathbb{P}\left(\overline{X} - c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}}) \leq \mu_* \leq \overline{X} + c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}})\right) \ = 1-\alpha,$$

which, despite the duality between hypothesis testing and CI, is not (9), but instead:

$$\frac{\sqrt{n}(\overline{X}-\mu_*)}{s} \overset{\text{TSN}}{\backsim} \mathsf{St}(n-1). \tag{10}$$

The difference in the underlying reasoning is important in understanding the nature of the error probabilities associated with each inference as well as in interpreting the results of these procedures.

The optimality of inference methods in frequentist statistics is defined in terms of their capacity to give rise to valid inferences (trustworthiness), evaluated in terms of the associated *error probabilities*: how often these procedures lead to erroneous inferences.

In the case of **Confidence Interval (CI) estimation** the capacity is usually assessed in terms of minimizing *the coverage error probability*:

$$\mathbb{P}\left(\overline{X} - c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}}) \leq \mu \leq \overline{X} + c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}});\ \mu \neq \mu_*\right) = \alpha,$$

the probability that the interval does *not* contain the true value μ_* of μ, or maximizing the *coverage probability*:

$$\mathbb{P}\left(\overline{X} - c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}}) \leq \mu \leq \overline{X} + c_{\frac{\alpha}{2}}(\tfrac{s}{\sqrt{n}});\ \mu = \mu_*\right) = 1-\alpha.$$

In the case of **hypothesis testing** the capacity of a test procedure is evaluated in terms of minimizing *the type II error probability*:

$$\mathbb{P}(\tau(\mathbf{X}) \leq c_\alpha; H_1(\mu_1)) = \beta(\mu_1), \text{ for all } \mu_1 > \mu_0,$$

the probability of accepting the null hypothesis when false, for a given *type I error (rejecting the null when true) probability*:

$$\mathbb{P}(\mathbf{x} \in C_1(\alpha); H_0) = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; H_0) = \alpha.$$

This is equivalent to maximizing the *power* of the test:

$$\pi(\mu_1) = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; H_1(\mu_1)) = 1-\beta(\mu_1), \text{ for all } \mu_1 > \mu_0;$$

see Lehmann (1986), Cox and Hinkley (1974).

Prediction differs from estimation (point and interval) and hypothesis testing in so far as it is concerned with *particular events* associated with the data generating process as opposed to the process itself. The predictor $\widehat{X}_{n+1} = h(\overline{X})$, where $h(.)$ is defined in (7), for a given data $\mathbf{x}_0$ gives rise to an *event* associated with the statistical GM $X_k = \mu + u_k,\ k \in \mathbb{N}$. The associated error is defined by $e_{n+1} = (X_{n+1} - \widehat{X}_{n+1})$, whose sampling distribution is directly related to (10). This suggests that the underlying reasoning in the case of prediction is also factual, as in the case of estimation.

2.6 Induction by enumeration

How do the above forms of statistical induction relate to the traditional *induction by enumeration* in philosophy of science? A particularly simple example of induction by enumeration is given by Salmon (1967):

"If the proportion of red marbles from a sample is (m/n), infer that approximately (m/n) of all marbles in the urn are red."

This can be given a precise statistical formulation in terms of the simple Bernoulli model where the outcome $X=1$ denotes the event 'the marble chosen is red', with $\mathbb{P}(X=1) = \theta$, and $X=0$ denotes the event 'the marble chosen is not red', with $\mathbb{P}(X=0) = 1-\theta$. This model differs from the simple Normal in table 1 only in so far as $E(X_k) = \theta$, $Var(X_k) = \theta(1-\theta)$ and the underlying distribution is Bernoulli, i.e.

$$X_k \backsim \mathsf{BerIID}\left(\theta,\ \theta(1-\theta)\right),\ k \in \mathbb{N}.$$

This formulation transforms the 'uniformity of nature' and 'representativeness' assumptions into the probabilistic assumptions of IID and renders them testable vis-a-vis data $\mathbf{x}_0$. Moreover, the inference concerning the proportion of red marbles amounts to choosing the point estimator $\widehat{\theta} = \overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$, as providing a representative value for θ; note that $\overline{x} = \left(\frac{m}{n}\right)$, m being the number of red marbles in a sample of n from an urn. The statistical justification of this form of induction in the early part of the 20th century was based on $\widehat{\theta}$ being a *consistent* estimator of θ as $n \to \infty$, under a variety of circumstances.

This formulation of induction by enumeration brings out some of its weaknesses most clearly. *First*, the inference in the form of point estimation is rather anemic without any measures of reliability. *Second*, reliance on consistency by itself:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| < \epsilon\right) = 1, \tag{11}$$

does not provide a basis for evaluating the reliability of inference because no trustworthy error probabilities can be retraced on the basis of (11). One needs to use

more effective inference procedures such as CI interval estimation or even testing, to evaluate the reliability of the inference concerning θ based on finite sample error probabilities such as:

$$\widehat{\theta} \backsim \mathsf{Ber}\left(\theta,\ \tfrac{\theta(1-\theta)}{n}\right).$$

Third, the premises of inference cannot be established a priori but can be empirically assessed a *posteriori*.

It is important to point out that the above explicit formulation of induction by enumeration, in the form of a simple Bernoulli model, is general enough to accommodate the modeling of the mechanism that brings about any event A, as long as the IID assumptions are appropriate. Moreover, it is crucial to distinguish between modeling the mechanism that can give rise to and predicting the occurrence of event A. For instance, predicting whether the next marble to be pulled out of the urn is red relates to the outcome associated with X_{n+1} whose best predictor is given in (7).

2.7 The role of pre-data error probabilities

In summary, the primary objective of statistical induction is to enable one to learn about the stochastic mechanism that gave rise to the data. This learning process begins with the choice of a statistical model (a family of models $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \boldsymbol{\theta} \in \Theta$) within the set of all possible statistical models $\mathcal{P}(\mathbf{x})$ that could have given rise to data $\mathbf{x}_0$. The selected statistical model, in conjunction with the data, constitute the premises of inference. An important necessary condition for the reliability of any inductive inference is the validity of the premises vis-a-vis data $\mathbf{x}_0$. Viewing the data as a truly typical realization of a generic observable stochastic process $\{X_k,\ k \in \mathbb{N}\}$ enables one to specify the premises in terms of testable probabilistic assumptions underlying this process. This is an important component of the objectivity characterizing the frequentist approach to inference. Securing the statistical adequacy of the premises using probative misspecification tests ensures that the actual and nominal error probabilities are approximately equal. Pre-data the error probabilities associated with an inductive inference provide a measure of the trustworthiness of an inference procedure: how often a certain procedure will give rise to valid inferences concerning the underlying actual DGM. As such ascertainable error probabilities, based on statistically adequate models, play a crucial role in statistical induction by determining the capacity/trustworthiness of the inference procedure under all possible sample realizations $\mathbf{x} := (x_1, ..., x_n) \in \mathcal{X}$, always within the prespecified family of models $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x}),\ \boldsymbol{\theta} \in \Theta$. That is the reason why all sampling distributions of estimators and test statistics are derived from the distribution of the sample $f(\mathbf{x}; \boldsymbol{\theta})$ via (3); this defines the domain of potential evidence. This role of pre-data error probabilities is well understood and generally accepted.

What is often disputed is the role of error probabilities *post-data*; see Hacking (1965). That is, once an inference is made what is the role of error probabilities, if

any, post-data? Are error probabilities inextricably bound up with the frequentist 'long-run' metaphor?

It is long-familiar that one cannot attach the coverage probability, say .95, to the observed CI (see Cox and Hinkley, 1974), and thus all points inside such an interval are treated on par post-data. Similarly, using pre-data error probabilities to go from the accept/reject decision to inferring the validity of the null or the alternative hypotheses often gives rise to the well-known fallacies of acceptance and rejection, respectively; see Mayo and Spanos (2006). The use of the well-known 'long-run' metaphor in interpreting such error probabilities has encouraged the view that error probabilities are only useful pre-data and just in the context of Neyman's behavioristic interpretation of tests. This issue lies at the heart of the tension between Fisher's significance testing and the Neyman-Pearson (N-P) hypothesis testing. Fisher's use of the *p-value* to provide an inferential interpretation reflecting the strength of evidence against the null, is often criticized as incompatible with N-P testing. Bayesians have admonished the usefulness of the p-value by pointing to the large n problem; they even dispute the interpretation of the p-value as a legitimate error probability (see Berger and Selke, 1987). They also correctly criticize the misconstrual of the p-value as assigning *a degree of support* or a *posterior probability* to the null hypothesis.

Despite these admonishments statistical practitioners have continued to use the p-value and when challenged they usually justify it by invoking some vague *inferential interpretation* based on the observed significance level. In several applied fields including Economics, Psychology, Epidemiology and Political Science, the behavioristic N-P accept/reject decision has been replaced by reporting a strange mixture of asterisks (significance at 1% (***), 5% (**), 10% (*)) and p-values. That, more than anything, suggests that practitioners are seeking ways to bridge the gap between the coarse accept/reject decision and the evidence for or against the null warranted by the data.

In the next section we will consider how an inferential construal of N-P tests can be attained by extending the pre-data error probabilities to a 'customized', post-data assessment of the severity with which specific inferences or claims pass the test in question; see Mayo (1991).

3 Statistical Induction : post-data

3.1 Error probabilities and an inferential construal of N-P tests

The type I and II error probabilities:

$$\alpha = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; \mu_0), \qquad \beta(\mu_1) = \mathbb{P}(\tau(\mathbf{X}) \leq c_\alpha; \mu_1), \text{ for all } \mu_1 > \mu_0,$$

are by definition fundamentally *pre-data* notions because they include (i) the choice of a predesignated α, giving rise to (ii) the critical value c_α, and (iii) both types of errors are relevant.

Any attempt to construct a *post-data* assessment would need to amend these notions in view of data $\mathbf{x}_0$ and the N-P accept/reject result. Crudely put, post-data (i)* the relevant significance level is the p-value:

$$p(\mathbf{x}_0) = \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu_0),$$

(not α), and (ii)* the relevant threshold is $\tau(\mathbf{x}_0)$ (not c_α); both are now data-specific. Moreover, post-data (iii)* the notion of error itself is different in the sense that it could only concern unwarranted claims associated with the accept/reject result. To establish that one needs to investigate the post-data 'trustworthiness' of the test procedure vis-a-vis such claims. The reasoning underlying this move is basically that of *learning from errors* by applying highly probative procedures which would have detected the error if it were present with very high probability.

In the same vain, data $\mathbf{x}_0$ provides evidence for a claim or a hypothesis H (H_0 or H_1) by applying a highly probative test which would have ruled out the ways the claim that H is correct can be in error, and no such error is detected. This intuition is formalized using the notion of a severe test; see Mayo (1981, 1996). A hypothesis H passes a *severe test* T with data $\mathbf{x}_0$ if,

(S-1) $\mathbf{x}_0$ agrees with H, and

(S-2) with very high probability, test T would have produced a result that accords less well with H than $\mathbf{x}_0$ does, if H were false.

This can be used to bridge the gap between accept/reject and an inferential interpretation in so far as the result that H passes test T provides good evidence for inferring H (is correct) to the extent that T severely passes H with data $\mathbf{x}_0$. By evaluating the severity of a test T, as it relates to claim H and data $\mathbf{x}_0$, we learn about the kind and extent of errors that procedures were (and were not) highly capable of detecting, thus informing us of errors ruled out and errors still in need of probing. Thus, from the thesis of learning from error, it follows that a severity assessment allows one to determine whether there is evidence for (or against) claims.

In statistical modeling a point null hypothesis is never *exactly true,* and when it's *false* one would like to 'quantify' the discrepancy from the null. With that in mind, the evaluation of severity introduces a *discrepancy parameter*. In the case of the *hypotheses* in the context of the simple Normal model:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0. \tag{12}$$

the discrepancy parameter γ is introduced via:

$$\mu_1 = \mu_0 + \gamma, \quad \text{for } \gamma \geq 0,$$

which is used to define the relevant *inferential claims* associated with H_0 and H_1.

In the case where $T_\alpha := \{\tau(\mathbf{X}), C_1(\alpha)\}$ has *rejected* H_0 the *relevant inferential claim* is:

$$\mu > \mu_1 = \mu_0 + \gamma, \ \gamma \geq 0,$$

and the idea is to establish the *largest discrepancy* γ from H_0 licensed by data $\mathbf{x}_0$.

In the case where T_α has *accepted* H_0 the *relevant inferential claim* is:

$$\mu \leq \mu_1 = \mu_0 + \gamma, \ \gamma \geq 0,$$

and the idea is to establish the *smallest discrepancy* γ from H_0 licensed by data $\mathbf{x}_0$.

In this sense the severity evaluation has three arguments, the test T_α, data $\mathbf{x}_0$ and a claim relating to the inference licensed by the data.

Case A. Reject H_0

(S-1) takes the form: $\mathbf{x}_0$ agrees with H_1,

(S-2) 'a result that accords less well with H_1 than $\mathbf{x}_0$ does'

can be formally written as $\{\mathbf{x} : \tau(\mathbf{x}) \leq \tau(\mathbf{x}_0)\}$.

Hence, in the case where test T_α resulted in *rejecting* H_0 (pass H_1) with data $\mathbf{x}_0$, the severity evaluation of the relevant claim $\mu > \mu_1$ is:

$$\mathsf{Sev}(T_\alpha; \mathbf{x}_0; \mu > \mu_1) := \mathbb{P}(\text{a result less in accord with } H_1 \text{ than } \tau(\mathbf{x}_0);\ H_1 \text{ is false}) =$$

$$= \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu > \mu_1 \text{ is false}) = \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_1).$$

Example. To illustrate consider the case of the simple Normal model with σ=2 and the hypotheses in (6). Let $\alpha = .025$, n=100, c_α=1.96. For $\tau(\mathbf{x}_0)$=4.0 ($\overline{x}$=12.8), then the severity of the claim $\mu > \mu_1$=12.4 is:

$$\text{(a)} \quad \mathsf{Sev}(T_\alpha; \mathbf{x}_0; \mu > 12.4) = \mathbb{P}(\tau(\mathbf{X}) \leq 4; \mu_1{=}12.4) = .977.$$

On the other hand, for $\tau(\mathbf{x}_0)$=2.0 ($\overline{x}$=12.4), then the severity of the claim $\mu > \mu_1$=12.4 is:

$$\text{(b)} \quad \mathsf{Sev}(T_\alpha; \mathbf{x}_0; \mu > 12.4) = \mathbb{P}(\tau(\mathbf{X}) \leq 2; \mu_1{=}12.4) = .5.$$

Our intuition here is that in case (b) one is not warranted in inferring so large a discrepancy ($\mu > 12.4$) on the basis of $\tau(\mathbf{x}_0)$=2.0, because 50% of the time an outcome as large as this would occur even if μ were no larger than 12.4. In case (a), however, such an inference is warranted on the basis of $\tau(\mathbf{x}_0) = 4.0$ because the severity is very high. Hence, two outcomes that might lead to an identical N-P decision, say, 'reject H_0 with a size α test' may license different inferences according to how severely the given rejection of H_0 indicates a discrepancy γ.

Case B. Accept H_0

(S-1) takes the form: $\mathbf{x}_0$ agrees with H_0,

(S-2) 'a result that accords less well with H_0 than $\mathbf{x}_0$ does'

can be formalized by $\{\mathbf{x} : \tau(\mathbf{x}) > \tau(\mathbf{x}_0)\}$.

Hence, in the case where test T_α resulted in *accepting* H_0 with data $\mathbf{x}_0$, the severity evaluation of the relevant claim $\mu \leq \mu_1$ is:

$$\mathsf{Sev}(T_\alpha; \mathbf{x}_0; \mu \leq \mu_1) := \mathbb{P}(\text{a result less in accord with } H_0 \text{ than } \tau(\mathbf{x}_0);\ H_0 \text{ is false}) =$$

$$= \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu \leq \mu_1 \text{ is false}) = \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu > \mu_1).$$

Example. Continuing the same example consider the case where T_α yields 'Accept H_0' with $\tau(\mathbf{x}_0) = 1.5$ ($\overline{x} = 12.3$). The severity of inferring $\mu \leq 12.3$:

$$\text{(a)} \quad \mathsf{Sev}(T_\alpha; \mathbf{x}_0; \mu \leq 12.3) = \mathbb{P}(\tau(\mathbf{X}) > 1.5; \mu > 12.3) = .500,$$

which indicates that no such claim is licensed by this data. On the other hand if the sample realization yielded $\tau(\mathbf{x}_0) = -0.5$ ($\overline{x} = 11.9$) :

$$\text{(b)} \quad \mathsf{Sev}(T_\alpha; \mathbf{x}_0; \mu \leq 12.3) = \mathbb{P}(\tau(\mathbf{X}) > -.5; \mu > 12.3) = .977.$$

For detailed discussions concerning severity vs. power, the use of severity to circumvent the fallacies of acceptance and rejection and the large n problem, as well as using severity to ameliorate the main shortcomings of the behavioral decision model of N-P tests, see Mayo and Spanos (2006).

3.2 Severity and post-data error probabilities

The question that naturally arises at this stage is the extent to which the severity evaluation can be viewed as a proper post-data error probabilistic assessment that addresses Fisher's concerns with the behavioristic construal of N-P tests and renders the inference data-specific. The simple answer is that the severity evaluation is a genuine post-data error probability which custom-tailors (in view of the data) the pre-data trustworthiness of a test procedure in order to weave an inferential interpretation out of the coarse accept/reject result and the relevant inferences warranted by data $\mathbf{x}_0$. Let us unpack this claims by foregrounding the custom-tailoring on both the sample and the parameter spaces.

Sample space custom-tailoring. N-P testing defines the pre-data error probabilities based on partitioning the sample space into an acceptance and a rejection region based on c_α :

$$C_0(\alpha) = \{\mathbf{x} : \tau(\mathbf{x}) \leq c_\alpha\}, \quad C_1(\alpha) = \{\mathbf{x} : \tau(\mathbf{x}) > c_\alpha\}, \quad C_0(\alpha) \cup C_1(\alpha) = \mathcal{X}.$$

The pre-data trustworthiness, in the sense of how often the test gives rise to the *correct decision,* is defined by:

$$\begin{array}{ll} \textbf{Accept } H_0: & \mathbb{P}(\text{accept } H_0 \text{ when true}) = 1 - \mathbb{P}(\text{reject } H_0 \text{ when true}) = \\ & = \mathbb{P}(\tau(\mathbf{X}) \leq c_\alpha; H_0) = 1 - \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; H_0) = 1-\alpha \\ \textbf{Reject } H_0: & \mathbb{P}(\text{reject } H_0 \text{ when false}) = 1 - \mathbb{P}(\text{accept } H_0 \text{ when true}) \\ & = \mathbb{P}(\tau(\mathbf{X}) > c_\alpha; H_1(\mu_1)) = 1 - \mathbb{P}(\tau(\mathbf{X}) \leq c_\alpha; H_1(\mu_1)) = 1-\beta(\mu_1), \end{array} \tag{13}$$

where α and $\beta(\mu_1)$, $\mu_1 > \mu_0$ denote the type I and II errors, respectively. Notice that the trustworthiness in the case of accept H_0 and reject H_0 is defined in terms of the type I and type II error, respectively.

Post-data, severity *re-partitions* the sample space into an *accordance* and *discordance regions* based on the threshold $\tau(\mathbf{x}_0)$:

$$A_0(\mathbf{x}_0) = \{\mathbf{x} : \tau(\mathbf{x}) \leq \tau(\mathbf{x}_0)\}, \quad A_1(\mathbf{x}_0) = \{\mathbf{x} : \tau(\mathbf{x}) > \tau(\mathbf{x}_0)\}, \quad A_0(\mathbf{x}_0) \cup A_1(\mathbf{x}_0) = \mathcal{X}.$$

These regions formalize the severity notions of 'accords as well as or better with H_0 than $\mathbf{x}_0$ does' and 'accords less well with H_0 than $\mathbf{x}_0$ does'; the reverse will be true for H_1.

Parameter space custom-tailoring. The second part of custom-tailoring involves the partitioning of the parameter space Θ using a partial order defined by

$$\mu_1 = \mu_0 + \gamma, \quad \text{for all } \gamma \geq 0,$$

with a fixed *reference point* $\mu = \mu_0$ and a varying discrepancy parameter $\gamma \geq 0$.

Collecting the above pieces together, the severity evaluation custom-tailors the pre-data trustworthiness to a post-data evidence-based assessment of the *relevant inferences* as follows:

$$\begin{aligned}
\mathsf{Sev}(T_\alpha;\mathbf{x}_0;\mu \leq \mu_1) \quad &= \mathbb{P}(\text{infer } \mu \leq \mu_1 \text{ when warranted by } \mathbf{x}_0) \\
&= \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu \leq \mu_1) \\
&\qquad \text{for } \mu_1 = \mu_0 + \gamma, \quad \gamma \geq 0, \qquad (14) \\
\mathsf{Sev}(T_\alpha;\mathbf{x}_0;\mu > \mu_1) \quad &= \mathbb{P}(\text{infer } \mu > \mu_1 \text{ when warranted by } \mathbf{x}_0) \\
&= \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu > \mu_1)
\end{aligned}$$

To show how the severity evaluation is directly related to post-data error probabilities associated with inferential claims, one can re-write (14):

$$\begin{aligned}
\mathsf{Sev}(T_\alpha;\mathbf{x}_0;\mu \leq \mu_1) \quad &= 1 - \mathbb{P}(\text{infer } \mu > \mu_1 \text{ when unwarranted by } \mathbf{x}_0) \\
&= 1 - \mathbb{P}(\tau(\mathbf{X}) > \tau(\mathbf{x}_0); \mu \leq \mu_1) > (1-\alpha) \\
&\qquad \text{for } \mu_1 = \mu_0 + \gamma, \quad \gamma \geq 0, \qquad (15) \\
\mathsf{Sev}(T_\alpha;\mathbf{x}_0;\mu > \mu_1) \quad &= 1 - \mathbb{P}(\text{infer } \mu \leq \mu_1 \text{ when unwarranted by } \mathbf{x}_0) \\
&= 1 - \mathbb{P}(\tau(\mathbf{X}) \leq \tau(\mathbf{x}_0); \mu > \mu_1) > [1-\beta(\mu_1)]\,.
\end{aligned}$$

A comparison between (15) and the pre-data trustworthiness in (13) indicates that the latter provide lower bounds (worst case scenarios) for the severity evaluations, highlighting the importance of custom-tailoring in rendering the inference more informative as well as data-specific.

The discrepancy parameter γ plays a crucial role in the severity assessment because it reflects the custom-tailored post-data trustworthiness of the test as it relates to different claims associated with the original accept/reject result. In a sense the counterfactual reasoning underlying N-P testing has been extended to cover the whole of the parameter space using the discrepancy-based partitioning as it relates to the relevant inference. In an important sense this intensive use of counterfactual reasoning provides the key to the data-specific inferential interpretation. Indeed, this explains why Confidence Interval (CI) estimation, based on factual reasoning, cannot discriminate among the points inside an observed CI.

Viewed from the severity perspective the p-value can be interpreted as a crude post-data error probability that lacks the discrepancy parameter refinement. To see this let us consider a severe-testing interpretation of using a small p-value, say $p = .01$, to infer that data $\mathbf{x}_0$ provide evidence against H_0.

Severe-testing and p-value. Such a small p-value indicates that $\mathbf{x}_0$ *accords with* H_1, and the question is whether it provides evidence for H_1. Using the severe-testing interpretation one can argue that H_1 has passed a severe test because the probability that test T_α would have produced a result that accords less well with H_1 than $\mathbf{x}_0$ does (values of $\tau(\mathbf{x})$ less than $\tau(\mathbf{x}_0)$), if H_1 were false (H_0 true):

$$\mathsf{Sev}(T_\alpha;\mathbf{x}_0;\mu>\mu_0)=\mathbb{P}(\tau(\mathbf{X})\leq\tau(\mathbf{x}_0);\mu\leq\mu_0) \quad =1-\mathbb{P}(\tau(\mathbf{X})>\tau(\mathbf{x}_0);\mu=\mu_0)=.99,$$

is very high. The severity construal of the p-value brings out its most crucial weakness: it establishes the existence of *some* discrepancy $\gamma\geq 0$, but provides no information concerning the magnitude licensed by data $\mathbf{x}_0$; the warranted discrepancy could be tiny or very large. Moreover, the dependence of the p-value on the sample size can belie the warranted discrepancy. The severity evaluation addresses both of these problems (Mayo and Spanos, 2006). This account is related to the principle for evidential interpretation governing the implications of p-values in Mayo and Cox (2006).

In summary, the severity-based inferential interpretation takes the form of rendering the coarse accept/reject decision more informative as well as data-specific by evaluating the discrepancy from the null that is licensed by the data in question. The question, however, remains whether the severity-based inferential interpretation addresses Fisher's concerns. How does the severity assessment relate to the metaphor of the 'long run' repetition of experiments used to conceptualize the pre-data error probabilities? The answer is that severity takes the pre-data error probabilities as calibrating the general trustworthiness of the test procedure and custom-tailors that to the particular case of data $\mathbf{x}_0$ and the relevant inferential claim, rendering the evaluation: (i) test-specific, (ii) data-specific, (iii) inference-based, (iv) discrepancy-driven, and (v) counterfactually-intensive.

The severity assessment allows for a post-data objective interpretation of any N-P test result that bridges the gap between the coarse accept/reject decision and the evidence for or against the null warranted by the data. When the severity evaluation of a particular inferential claim, say $\mu\leq\mu_0+\gamma$, is very high (close to one), it can be interpreted as indicating that this claim is warranted to the extent that the test has ruled out discrepancies larger than γ; the underlying test would have detected a departure from the null as large as γ almost surely, and the fact that it didn't suggests that no such departures were present. Viewing N-P tests from the severe testing perspective, suggests that the value of confining *error probabilities* at small values is not only the desire to have a good track record in the *long run*, but also because of how this lets us severely probe, and thereby learn about, the process that gave rise to the particular data. This brings us back to *learning from errors* by applying highly probative procedures.

In concluding this section, it is important to emphasize that the severity assessment constitutes a general post-data supplement to N-P tests that provides a data-specific inferential interpretation of the accept/reject result; it can be applied to all

(properly defined) N-P tests. In particular, it does not require that the underlying statistical model enjoys any special probabilistic structure such as the existence of *ancillary statistics,* as in the case of *conditional inference* (see Lehmann, 1986, ch. 10). Moreover, it does not require any special partitioning of the sample space arising from additional information that needs to be brought into the N-P framework, as in the case of Kiefer's (1977) *conditional confidence* set up.

4 Model validity and the reliability of Inference

As argued above, the sampling distributions, in terms of which the pre-data and post-data error probabilities are evaluated, are derived from the distribution of the sample. The distribution of the sample is the joint distribution of the stochastic process defined by the statistical model; the premises of inference. Hence, it follows that the reliability of any inductive inference depends crucially on *statistical adequacy*: the model assumptions are valid vis-a-vis the data in question. The Fisher-Neyman probabilistic perspective is particularly crucial in this context because it enables one to specify the statistical model in terms of complete set of testable probabilistic assumptions. The quintessential example of a statistical model is the simple Normal model given in table 1.

Any departures from the model assumptions will give rise to a divergence between the *nominal* error probabilities, derived under the assumption of valid premises, and the *actual* error probabilities, derived taking into consideration the particular departure(s) from the premises, calling into question the reliability of inference. Indeed, the discrepancy between the nominal and actual error probabilities provides a way to assess the extent of the unreliability of inference. Because of that, statistical adequacy is viewed as operationally equivalent to the condition that the nominal and actual error probabilities are approximately equal, giving rise to statistically reliable inferences.

4.1 Misspecification and the reliability of inference

The question which arises is how does one deal with departures from the statistical premises? A widely used argument in defence of ignoring the problem of potential misspecifications is the following:

> "All models are misspecified to 'a greater or lesser extent' because they are just approximations. Moreover, 'slight' departures from the assumptions will only lead to 'minor' deviations from the 'optimal' inferences."

What is misleading about this argument is that its persuasive force stems from its vague references to 'lesser' and 'slight', and 'minor' which, when taken at face value, seem plausible until one examines this an argument more closely.

Example. Consider the case of the simple Normal model with σ^2 known ($\sigma^2 = 1$)

and the hypotheses of interest are:

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu > \mu_0, \tag{16}$$

When assumptions [1]-[4] are valid the t-type test based on the sampling distributions:

$$d(\mathbf{X})=\frac{\sqrt{n}(\overline{X}-\mu_0)}{\sigma} \overset{H_0}{\backsim} \mathsf{N}(0,1), \qquad d(\mathbf{X}) \overset{H_1}{\backsim} \mathsf{N}(\delta,1), \text{ for } \mu_1 > \mu_0, \tag{17}$$

where $\delta = \frac{\sqrt{n}(\mu_1-\mu_0)}{\sigma}$, together with the rejection region $C_1(\alpha) = \{\mathbf{x} : d(\mathbf{x}) > c_\alpha\}$ is UMP; see Lehmann (1986). As shown below, however, the presence of some dependence of the form:

$$Corr(X_i, X_j) = \rho, \ 0<\rho<1, \ i \neq j, \ i,j = 1,...n, \tag{18}$$

will render this test *unreliable.* The unreliability of inference arises when one applies this test thinking that there is only 5% chance of rejecting the null when true, when in fact that probability could be as high as 100%; see Spanos and McGuirk (2001) for several examples.

Let $\mu_0 = 0$, $n = 100$, $\alpha = .05$, $c_\alpha = 1.66$. Table 2 shows that the presence of even some tiny correlation ($\rho = .05$) will induce a sizeable discrepancy between the *nominal* ($\alpha = .05$) and *actual type I error probability* ($\alpha^* = .25$). In the above case the unreliability stems from the fact that actual the sampling distributions are no longer given by (17) but:

$$d(\mathbf{X}) \overset{\mu=\mu_0}{\backsim} \mathsf{N}\left(0, d_n(\rho)\right), \quad d(\mathbf{X}) \overset{H_1}{\backsim} \mathsf{N}(\delta, d_n(\rho)), \text{ for } \mu_1 > \mu_0, \tag{19}$$

where $d_n(\rho)=(1+(n-1)\rho) > 1$.

Table 2 - Type I error of t-test for different values of ρ

ρ	**0.0**	.05	.10	.20	.30	.50	.75	.80	.90
α^*-actual	**.05**	.249	.309	.359	.383	.408	.425	.427	.431

Table 3 - Power $\pi^*(\mu_1)$ of the t-test for different values of ρ

ρ	$\pi^*(.01)$	$\pi^*(.02)$	$\pi^*(.05)$	$\pi^*(.1)$	$\pi^*(.2)$	$\pi^*(.3)$	$\pi^*(.4)$
0.0	**.061**	**.074**	**.121**	**.258**	**.637**	**.911**	**.991**
.05	.262	.276	.318	.395	.557	.710	.832
.1	.319	.330	.364	.422	.542	.659	.762
.2	.367	.375	.401	.443	.531	.616	.697
.3	.390	.397	.418	.453	.525	.596	.664
.5	.414	.419	.436	.464	.520	.575	.630
.75	.429	.434	.447	.470	.516	.562	.607
.8	.431	.436	.449	.471	.515	.560	.603
.9	.435	.439	.452	.473	.514	.556	.598

Similarly affected will be the power of the t-test. As shown in table 3, as $\rho \to 1$ the power of the t-test increases for small discrepancies from the null, but it decreases for larger discrepancies. That is, the presence of correlation would render a powerful smoke alarm into a *faulty one,* being triggered by burning toast but not sounding until the house is fully ablaze; see Mayo (1996).

Misspecification renders CIs unreliable by inducing a discrepancy between nominal and actual coverage probabilities analogous to the ones for tests considered above. The two-sided $(1-2\alpha)$ CI takes the form:

$$\mathbb{P}\left(\overline{X} - c_\alpha(\tfrac{\sigma}{\sqrt{n}}) \leq \mu^* < \overline{X} + c_\alpha(\tfrac{\sigma}{\sqrt{n}})\right) = 1 - 2\alpha, \tag{20}$$

with length: $\frac{\sigma 2 c_\alpha}{\sqrt{n}}$.

Example. For $\overline{x} = 0.6,\ \alpha = .025,\ c_\alpha = 1.96,\ \sigma = 1$ and $n = 100$, the observed 95% CI is:

$$CI(\mathbf{x}_0) = [0.404,\ 0.796]\,, \tag{21}$$

of length 0.392. However, when assumption [4] is false, and instead (18) is the appropriate assumption, the actual sampling distribution of the relevant *pivotal quantity* is:

$$\frac{\sqrt{n}(\overline{X}-\mu^*)}{\sigma\sqrt{d_n(\rho)}} \overset{\text{TSN}}{\backsim} \mathsf{N}(0,1)\,. \tag{22}$$

Hence, the actual coverage probability of the CI becomes:

$$\mathbb{P}\left(\overline{X} - c_\alpha(\tfrac{\sigma\sqrt{d_n(\rho)}}{\sqrt{n}}) \leq \mu^* < \overline{X} + c_\alpha(\tfrac{\sigma\sqrt{d_n(\rho)}}{\sqrt{n}})\right) = 1 - 2\alpha^*, \tag{23}$$

of length: $\frac{\sigma 2 c_\alpha \sqrt{d_n(\rho)}}{\sqrt{n}}$. As shown in table 4, the presence of the misspecification induces a discrepancy, not only between nominal and actual coverage probabilities, but also between the nominal and actual lengths.

Table 4 - CIs under Misspecification

ρ	Actual CI	Actual cover. prob.
0.0	**[0.404, 0.796]**	**.950**
.05	[0.122, 1.078]	.578
.1	[-.047, 1.247]	.447
.2	[-.294, 1.494]	.333
.3	[-.486, 1.686]	.276
.5	[-0.793, 1.993]	.217
.75	[-1.100, 2.300]	.179
.8	[-1.155, 2.355]	.173
.9	[-1.260, 2.460]	.163

Given the potential serious consequences of statistical misspecifications, how does one deal with the problem. First, one needs effective procedures which will detect the

presence of misspecifications. This is the subject matter of misspecification testing. Second, if any departures are detected, one needs a respecification strategy which will lead to a statistical adequate model.

4.2 Misspecification Testing

MisSpecification (M-S) testing, understood as assessing the validity of the statistical premises, raises several important methodological problems; see Mayo and Spanos (2004). Some of these problems stem from the difference in nature between N-P and M-S testing. In broad terms, N-P testing assumes that the prespecified statistical model $\mathcal{M}$ includes the true model, say $f_0(\mathbf{x})$, and probes *within the boundaries* of this model class by partitioning it into two subsets, the null and alternative:

$$H_0 : f_0(\mathbf{x}) \in \mathcal{M}_0 \text{ vs. } \ H_1 : f_0(\mathbf{x}) \in \mathcal{M}_1,$$

where $\mathcal{M}_0$ and $\mathcal{M}_1$ form a partition of $\mathcal{M}$. In contrast M-S testing probes *outside the boundaries* of the prespecified model:

$$H_0 : f_0(\mathbf{x}) \in \mathcal{M} \text{ vs. } \overline{H}_0 : f_0(\mathbf{x}) \in [\mathcal{P}-\mathcal{M}]\,,$$

where $\mathcal{P}$ denotes the set of all possible statistical models that can be specified in terms of the joint distribution $D(X_1, X_2, ..., X_n; \boldsymbol{\phi})$. The problem with M-S testing is how one can operationalize $\mathcal{P}-\mathcal{M}$ in order to probe thoroughly for possible departures. In view of the fact that $\overline{H}_0$ can take a (possibly) infinite number of forms, deriving a test requires the modeler to provide a more restrictive (operational) form, $\mathcal{P}_1 \subset [\mathcal{P}-\mathcal{M}]$, where $\mathcal{P}_1$ can be as vague as a direction of departure from $\mathcal{M}$ or as specific as a proper statistical model that encompasses $\mathcal{M}$ ($\mathcal{M} \subset \mathcal{P}_1$). For an extensive discussion of M-S testing in relation to the simple Normal model see Spanos (1999), ch. 15.

The severe testing reasoning is particularly useful in this context because M-S testing is more susceptible to the fallacies of acceptance and rejection than N-P testing because the stated alternative $\mathcal{P}_1$ can take many different forms, none of which need to constitute an appropriate statistical model. In particular, inferring that a specific misspecification error is ruled out, when the test had no chance to detect such a departure, will be unwise but all too easy. Similarly, rejecting the null in a M-S test does not warrant one to infer the validity of the specified alternative $\mathcal{P}_1$. Detection of departures from $\mathcal{M}$ in the direction of $\mathcal{P}_1$ is sufficient to consider the null as false but not to consider $\mathcal{P}_1$ as true. In both cases $\mathcal{M}$ and $\mathcal{P}_1$, respectively *have not passed a severe test*; see Spanos (2000), Mayo and Spanos (2004).

To guard against such fallacies the Probabilistic Reduction approach encourages an exhaustively complete probing strategy for *M-S testing*, by using:

(i) the reduction assumptions to guide the probing in directions of potential departures,

(ii) graphical techniques for *informed probing*,

(iii) a judicious combination of ordered parametric and non-parametric tests to *avoid circular reasoning*, and

(iv) joint M-S tests (testing several assumptions simultaneously) to avoid erroneous diagnoses.

The challenge is to arrive at a statistical model for the process underlying the data and *infer with severity* (Mayo, 1996) that potential violations of its assumptions have been well-probed! In Spanos (2005b) it is shown that one can do this effectively using joint M-S tests based on artificial regressions constructed by employing orthogonal polynomials; see empirical examples in Spanos (1986, 1995, 2006) and Spanos and McGuirk (2001).

4.2.1 M-S testing and double-use of data

Worrall (2002) defines use-novelty by 'you can't use evidence in the construction of a theory and then again in its support'. Could the use of graphical techniques in, specification and M-S testing be considered as an example of violating use-novelty or double use of data?

M-S testing raises the issue of double use of data in the sense that data $\mathbf{x}_0$ is initially used (i) to probe the statistical adequacy of the model, and then (ii) to test the primary hypotheses or claims. Does this constitute a pejorative double use of data? The answer to that question is twofold.

The methodological answer is that the questions posed to the data in (i) and (ii) are very different. In (i) one assesses the appropriateness of the statistical model by probing the claim that 'data $\mathbf{x}_0$ constitute a truly typical realization of the stochastic process $\{X_k,\ k\in\mathbb{N}\}$ underlying this statistical model'. Hence, it concerns this particular data $\mathbf{x}_0$ vis-a-vis the selected statistical model. In contrast, (ii) assesses a claim about the underlying data-generating process itself. When one tests the hypothesis $\mu = 12$, in the context of the simple Normal model with σ known, one is asking the question whether:

$$X_k = 12 + \sigma\varepsilon_k,\ \varepsilon_k \backsim \mathsf{N}(0,1),\ k\in\mathbb{N},$$

captures (statistically) the true data-generating process; see Spanos (2000).

There is also a related formal statistical answer which concerns how (i) and (ii) depend on different information, rendering them unrelated. It can be shown that in the case of the simple Normal model ($\mathcal{M}$), one can reduce the distribution of the sample into two unrelated components:

$$f(\mathbf{x};\boldsymbol{\theta}) = |J|\cdot f(\mathbf{s};\boldsymbol{\theta})\cdot f(\mathbf{r}) \text{ for all } (\mathbf{r},\mathbf{s})\in\mathbf{R}_X^n, \tag{24}$$

where $\mathbf{s} = (\overline{X}, s)$:

$$\overline{X} = \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i, \qquad s^2 = \tfrac{1}{n(n-1)}\textstyle\sum_{i=1}^n (X_i-\overline{X})^2,$$

constitute *minimal sufficient statistics* for $\boldsymbol{\theta} = (\mu,\sigma^2)$ and $\mathbf{r} := (\widehat{v}_3,..,\widehat{v}_n)$:

$$\widehat{v}_k = \tfrac{(X_k-\overline{X})}{s}, \qquad k = 3,4,..,n,$$

are *maximal ancillary statistics*. This defines a one-to-one transformation:

$$(X_1, X_2, ..., X_n) \longleftrightarrow (\overline{X}, s, \widehat{v}_3, .., \widehat{v}_n),$$

with $|J|$ the Jacobian of this transformation. When these conditions hold one can argue formally that (a) any inference concerning $\boldsymbol{\theta}$ should be based solely on $f(\mathbf{s};\boldsymbol{\theta})$ because it's the only factor that involves $\boldsymbol{\theta}$ in a way which ensures no loss of information, and (b) since $f(\mathbf{r})$ is free of the unknown parameters $\boldsymbol{\theta}$, it can be used to assess the statistical adequacy of model $\mathcal{M}$; see Spanos (2006) for the details.

A related issue raised by both specification and M-S testing is that the observed data, depicted in terms of graphical techniques, play an important role at these facets of modeling. After all, the specified statistical model $\mathcal{M}$ is so chosen so as to render data $\mathbf{x}_0$ a truly typical realization of the stochastic process $\{X_k,\ k\in\mathbb{N}\}$ underlying $\mathcal{M}$. Graphs of the data indicate recurring patterns that reflect the structure of the underlying stochastic process, and thus studying these graphs can contribute significantly to the 'economy of thought' required to choose an appropriate statistical model.

4.3 Respecification

What happens if some of the model assumptions are found wanting? *Respecify*: select another statistical model whose appropriateness will be assessed by how well it can account for the systematic statistical information the original model could not.

In the context of the PR approach, respecification (theoretically) takes the form of tracing the results of the misspecification tests back to the reduction assumptions and then changing the reduction assumptions judiciously to account for the information in the detected departures in order to specify a more appropriate statistical model.

To illustrate this, let us return to the specification of the simple Normal model as given in table 1. As argued above, the PR perspective considers this model as a reduction of the joint distribution of the process $\{X_k\ k\in\mathbb{N}\}$ by imposing the probabilistic assumptions of NIID:

$$D(X_1, X_2, ..., X_n;\boldsymbol{\varphi}) \overset{\mathsf{I}}{=} \prod_{k=1}^{n} D_k(X_k;\boldsymbol{\phi}_k) \overset{\mathsf{IID}}{=} \prod_{k=1}^{n} D(X_k;\boldsymbol{\phi}). \tag{25}$$

Consider the case where assumption [4] (see table 1) is false, and instead the sample is *Markov dependent:*

$$\boxed{\text{[5] } Corr(X_i, X_j) = \rho^{|i-j|},\ \text{for } -1<\rho<1,\ \text{for all } i\neq j,\ i,j=1,...n.} \tag{26}$$

This departure suggests that the simple Normal model is no longer appropriate and the question of respecification arises. As argued in Spanos (1999), ch. 15, under (26) the appropriate statistical model suggested by the PR approach comes in the form of the *Autoregressive (AR(1)) model,* as specified in table 5.

The probabilistic reduction in (25) is no longer appropriate, and by replacing Independence (I) with Markov (M) dependence and extending Identically Distributed (ID) to Stationarity (S), the appropriate reduction takes the form:

$$D(X_1, X_2, \cdots, X_n; \boldsymbol{\phi}) \overset{\mathsf{M}}{=} D_1(X_1; \boldsymbol{\psi}_1) \prod_{k=2}^{n} D_k(X_k|X_{k-1}; \boldsymbol{\psi}_k) = \overset{\mathsf{M\&S}}{=} D_1(X_1; \boldsymbol{\psi}_1) \prod_{k=2}^{n} D(X_k|X_{k-1}; \boldsymbol{\psi}). \tag{27}$$

This gives rise to the Autoregressive (AR(1)) model (see table 5), where the *statistical parameterization* of the unknown parameters $(\alpha_0, \alpha_1, \sigma_0^2)$ is:

$$\begin{array}{l} \alpha_0 = E(X_t) - \alpha_1 E(X_{t-1}) = \mu(1-\alpha_1) \in \mathbb{R}, \quad \alpha_1 = \frac{Cov(X_t, X_{t-1})}{Var(X_{t-1})} = \frac{\sigma(1)}{\sigma(0)} = \rho \in (-1, 1), \\ \sigma_0^2 = \sigma(0) - \frac{[\sigma(1)]^2}{\sigma(0)} = \sigma(0)\,(1 - \alpha_1^2) \in \mathbb{R}_+. \end{array} \tag{28}$$

Table 5 - Normal AutoRegressive Model

Statistical GM:		$X_k = \alpha_0 + \alpha_1 X_{k-1} + \varepsilon_t,\ k \in \mathbb{N}.$	
[1]	Normality:	$(X_k \mid X_{k-1}) \backsim \mathsf{N}(.,.),$	$k \in \mathbb{N}.$
[2]	Linearity:	$E\,(X_k \mid X_{k-1}) = \alpha_0 + \alpha_1 X_{k-1},$	
[3]	Homoskedasticity:	$Var\,(X_k \mid X_{k-1}) = \sigma_0^2,$	
[4]	Markov dependence:	$\{X_k,\ k \in \mathbb{N}\}$ is a Markov process,	
[5]	k-invariance:	$(\alpha_0, \alpha_1, \sigma_0^2)$ are *not* changing with k,	

(29)

(28) brings out the relationship between the AR(1) parameters $(\alpha_0, \alpha_1, \sigma_0^2)$ and the parameters of the simple Normal model (μ, σ^2); see Spanos (1999). Note that the presence of μ, as part of the implicit parametrization of (α_0, α_1), enables one to test the hypotheses (16) in the context of the AR(1) model (29). This can be done in the context of the reparameterized model:

$$\Delta X_t = \gamma_0 + \beta_1\,(X_{t-1} - \mu_0) + u_t, \quad t \in \mathbb{T}, \tag{30}$$

$$H_0 : \gamma_0 = 0 \quad \text{vs.} \quad H_1 : \gamma_0 > 0. \tag{31}$$

see Spanos (2005a) for the details.

4.4 Model specification vs. model selection

The problem of statistical model specification is often conflated with that of model selection, as currently understood in the statistical literature; see Rao and Wu (2001). Lehmann (1990) pointed out that the current *model selection* procedures do not address the statistical model specification problem; indeed, one can make a strong case that they assume the latter problem solved. The model selection procedure is applied

to a particular family of parametric models which is assumed to be overparameterized, but contains the 'true' model. Consider the classic example of using a model selection procedure to choose the 'optimal' lag value p in the AR(p) model (table 6).

The problem addressed by such a procedure is to choose $p \geq 1$ on the basis of some criterion, say Akaike's information criterion. This, however, presupposes (implicitly) that assumptions [1]-[5] comprising the AR(p) family of models are valid for data $\mathbf{x} := (x_1, x_2, ..., x_n)$, and the only issue that remains is the choice of p. When any of the assumptions [1]-[5] are invalid, however, the selection procedure is likely to lead one astray because, both, the likelihood function and any error probabilities utilized, are likely to be misleading. Indeed, this argument can be used to make a strong case for Glymour's (1981) position that 'goodness of fit' traded against 'simplicity', does not provide an adequate procedure to choose the fittest model.

Table 6 - Normal AutoRegressive (AR(p)) Model

Statistical GM:		$y_k = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{k-i} + u_k, \quad k \in \mathbb{N}.$	
[1]	Normality:	$\left(y_k \middle\vert \mathbf{y}_{k-1}^0\right) \backsim \mathsf{N}(.,.),\ \mathbf{y}_{k-1}^0 := (y_{k-1}, ..., y_1)$	$\left.\begin{array}{l}\\\\\\\\\end{array}\right\}$ $k \in \mathbb{N}.$
[2]	Linearity:	$E\left(y_k \middle\vert \mathbf{y}_{k-1}^0\right) = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{k-i},$	
[3]	Homoskedasticity:	$Var\left(y_k \middle\vert \mathbf{y}_{k-1}^0\right) = \sigma_0^2,$	
[4]	Markov (p):	$\{y_k,\ k \in \mathbb{N}\}$ is a Markov (p) process,	
[5]	k-invariance:	$(\alpha_0, \alpha_1, \sigma_0^2)$ are *not* changing with k,	

(32)

In addition, when the statistical adequacy issue is addressed at the specification stage, the problem solved by a model selection procedure becomes superfluous. Consider the case where the statistical model specification problem is addressed using thorough M-S testing and respecification to achieve statistical adequacy. Part of establishing statistical adequacy is the choice of the maximum needed p in order to capture the order of Markov dependence of the underlying process process $\{X_k,\ k \in \mathbb{N}\}$; rendering the residuals non-systematic. This will solve the choice of p problem on statistical adequacy grounds, rendering the model selection procedure redundant; see Spanos (2006c) for further details.

The model selection methods based on information criteria, such as the AIC, would often lead to erroneous inferences even in cases where the true model is a member of the class of models chosen at the outset. The unreliability of these procedures stems from the fact that their minimization procedures can be shown to be equivalent to N-P testing but without controlling the error probabilities; see Spanos (2006c) for the details.

4.5 Statistical vs. Substantive adequacy

The relationship between statistical and substantive information in empirical modeling is highly embrangled (see Lehmann, 1990, Cox, 1990 and Cox and Warmuth, 1996), and no attempt will be made in this paper to disentangle the intricate connections; see Spanos (2006a-c). However, it is important to discuss the issue of statistical vs. substantive adequacy which is an important dimension of that relationship.

Broadly speaking, *statistical adequacy* concerns the validity of the *statistical model* (the probabilistic assumptions constituting the model - table 1- vis-a-vis the observed data. As argued above, statistical adequacy ensures that the actual error probabilities provide a good approximation to the nominal error probabilities, rendering the inference based on such a model reliable. This, however, is not sufficient for substantive adequacy. *Substantive adequacy* concerns the validity of the *structural model* (the inclusion of relevant and the exclusion of irrelevant variables, functional relationships, confounding factors, causal claims, external validity, etc.) vis-a-vis the phenomenon of interest that gave rise to the data. The two premises are related in so far as the statistical model provides the operational context in which the structural model can be analyzed, but the nature of errors associated with the two premises is very different. Moreover, a structural model gains statistical 'operational meaning' when embedded into a statistically adequate model. In an attempt to illustrate some of these issues we revisit Kepler's first law using his original data.

4.5.1 Kepler's first law of planetary motion revisited

Historically this law was originally proposed by Kepler in 1609 as an *empirical regularity* (a statistical model) that he 'deduced' from Brahe's data. Almost 80 years later Newton proposed a substantive explanation (structural model) for this regularity. In what follows we will reverse this order for expositional purposes.

Structural Model. Kepler's first law states that the loci of the motion in *polar coordinates* can be approximated by $(1/r) = \alpha_0 + \alpha_1 \cos\vartheta$, where r denotes *the distance of the planet from the sun*, and ϑ denotes *the angle between the line joining the sun and the planet and the principal axis of the ellipse*; see figure 7.

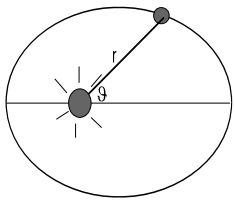

Fig. 7: Elliptical motion of planets

Defining the observable variables, $Y := (1/r)$ and $X := \cos\vartheta$, one can specify the

structural model:

$$Y_k = \alpha_0 + \alpha_1 X_k + \epsilon(x_k,\boldsymbol{\xi}_k),\ k \in \mathbb{N}, \tag{33}$$

where the error term $\epsilon(x_k,\boldsymbol{\xi}_k)$ includes all the unmodeled effects and is assumed to be white-noise.

The structural interpretation of Kepler's first law, as given in (33), stems from the fact that the parameters (α_0, α_1) enjoy a clear *theoretical interpretation*. This interpretation is bestowed upon the parameters by Newton's law of universal gravitation: $F = \frac{G(m \cdot M)}{r^2}$, where F is the force of attraction between two bodies of *mass* m (planet) and M (sun); G is a *constant of gravitational attraction,* and r is the *distance* between the two bodies. In particular, the parameters (α_0, α_1) are given the following clear *structural interpretation:*

$\alpha_0 = \frac{MG}{4\kappa^2}$, where κ denotes Kepler's constant,

$\alpha_1 = (\frac{1}{d} - \alpha_0)$, d denotes the shortest distance between the planet and the sun.

The error term $\epsilon(x_k,\boldsymbol{\xi}_k)$ also enjoys a structural interpretation in the form of 'discrepancies' from the elliptic motion due to *measurement errors* and other *unmodeled effects*. Hence, the white-noise error assumptions are *inappropriate* in cases where:

(i) the data suffer from 'systematic' observation errors,

(ii) the *third body problem* and/or the *general relativity* terms (see Lawden, 2002) are significant.

In summary, the structural model (33) has the following crucial features:

(a) it depicts a 'factual' generating mechanism which aims to approximate the actual Data Generating Mechanism, viewed as a 'nearly isolated' system where the unmodeled effects are non-systematic,

(b) the parameters $(\alpha_0, \alpha_1, \sigma_\epsilon^2)$ enjoy a clear substantive interpretation,

(c) the error term is 'autonomous' and represents all unmodeled influences.

Statistical model. In order to assess the substantive adequacy of the structural model (33) one needs to embed it into a statistical model. The obvious choice of a statistical model is the Linear Regression given in table 7. Viewed in the context of that model, Kepler's law constitutes an *empirical regularity* if the estimated model turns out to be statistically adequate.

Table 7 - The Normal/Linear Regression (LR) Model

Statistical GM:	$y_t = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t + u_t,\ t \in \mathbb{T},$
[1] Normality:	$(y_t \mid \mathbf{X}_t = \mathbf{x}_t) \backsim \mathsf{N}(.,.),$
[2] Linearity:	$E(y_t \mid \mathbf{X}_t = \mathbf{x}_t) = \beta_0 + \boldsymbol{\beta}_1^\top \mathbf{x}_t$, linear in $\mathbf{x}_t$,
[3] Homoskedasticity:	$Var(y_t \mid \mathbf{X}_t = \mathbf{x}_t) = \sigma^2$, free of $\mathbf{x}_t$,
[4] Independence:	$\{(y_t \mid \mathbf{X}_t = \mathbf{x}_t),\ t \in \mathbb{T}\}$ is an independent process,
[5] t-invariance:	$\boldsymbol{\theta} := (\beta_0, \boldsymbol{\beta}_1, \sigma^2)$ do not vary with t.
$\beta_0 = \mu_1 - \boldsymbol{\beta}_1^\top \boldsymbol{\mu}_2,$	$\boldsymbol{\beta}_1 = \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21},\ \sigma^2 = \sigma_{11} - \boldsymbol{\sigma}_{21}^\top \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\sigma}_{21}$

(34)

To assess that, consider embedding the structural model (33) into a statistical (Linear Regression) model (table 7):

$$y_t = \underset{(.000002)}{0.662062} + \underset{(.000003)}{.061333}x_t + \widehat{u}_t, \quad n = 28, \ s{=}.0000111479, \ R^2 = .9999, \tag{35}$$

where $\widehat{\alpha}_0 = .662062$ and $\widehat{\alpha}_1 = .061333$. The M-S test results reported in table 8 are indicative of the thorough probing of assumptions [1]-[5], which included both parametric and non-parametric tests. The numbers in square brackets denote the relevant p-values, which indicate no departures from the model assumptions [1]-[5]; see Spanos and McGuirk (2001) for the details of the M-S tests. Note that the residuals from (35) indicate no systematic departures from a realization of a NIID process.

Table 8 - Misspecification tests	
Non-Normality:	$D'AP = 5.816[.056]$
Non-linearity:	$F(1,25) = 0.077[.783]$
Heteroskedasticity:	$F(2,23) = 2.012[.156]$
Autocorrelation:	$F(2,22) = 2.034[.155]$

In view of these M-S test results we can deduce that any inference based on (35) will be statistically reliable. For instance, one can test Kepler's first law of motion being elliptical against Copernicus's conjecture that the motion was circular by testing:

$$H_0 : \alpha_1 = 0, \ \text{vs.} \ H_1 : \alpha_1 \neq 0.$$

This is because the equation of a circle in polar coordinates is $(1/r) = \alpha_0 \neq 0$.

Fig. 8: Residuals from the Kepler regression

Substantive adequacy. The question is whether this estimated model is also substantively adequate. To assess that one would need to modify the above structural model to include the *general relativity* factor and any measurement errors. This factor

arises from Einstein's General Relativity Theory induced 'correction' of Newton's law of universal gravitation to:

$$F^* = \frac{G(m \cdot M)}{r^2} + \frac{A}{r^4};$$

see Lawden (2002). The general relativity factors turns out to be tiny, but the potential of systematic measurement errors committed by Brahe remains a real possibility. To assess this we use a variable known as the *Julian date* which can play the role of a proxy for any omitted effects relating to the observation order, as a potentially omitted variable; see Spanos (2006b). The inclusion of ξ_k is suggested by the t-plots of the data which exhibit clear trends. Re-estimating (35), with the additional variables yielded:

$$y_t = \underset{(.000005)}{.66206} + \underset{(.000005)}{.061328}x_t + \underset{(.00226)}{.00278}\xi_k + \widehat{u}_t, \quad n = 28, \ s{=}.000011023, \ R^2 = .9999. \tag{36}$$

Thorough misspecification testing of assumptions [1]-[5], reveals that (36) is indeed statistically adequate; the results are very similar to those in table 8. The fact that Kepler's empirical relationship turned out to constitute a statistically adequate model was primarily due to luck; a combination of accurate observations from Brahe on Mars, and the long distance of Mars from the nearest planet which rendered the *third body problem* effect negligible.

The data-acceptability of the structural model (33) can be assessed on the basis of (36) by testing the statistical significance of β_2, yielding: $\tau(\mathbf{y}) = 1.230[.231]$, i.e. the null hypothesis cannot be rejected. In this sense, the estimated structural model (35) constitutes a *reparameterization/restriction* of the estimated statistical model (36). Despite the small sample size the precision of the structural estimates, based on their standard errors, is amazing!

It is interesting to note that on the basis of the estimates of (α_0, α_1) in (35), one can proceed to derive indirect estimates of several other related structural parameters:

$$\widehat{\gamma}_0 = \tfrac{1}{\widehat{\alpha}_0} = 1.51043, \quad \widehat{\gamma}_1 = \widehat{\alpha}_1\widehat{\gamma}_0 = .0926392, \quad d = \tfrac{1}{\widehat{\alpha}_0+\widehat{\alpha}_1} = 1.3824,$$

where γ_0 denotes the *semi-latus rectum,* γ_1 the *eccentricity* of the elliptical motion of Mars around the sun. These indirect estimates are surprisingly very accurate, even when compared with current estimates; see Spanos (2005d) for further details.

5 Summary and conclusions

A number of methodological problems and issues have been discussed using the Probabilistic Reduction approach to empirical modeling. A central axis around which most of these issues revolve is the notion of statistical induction adopted. The paper articulated the notion of statistical induction underlying the frequentist approach going back to Fisher (1922), emphasizing the fact that the primary objective is to learn

about the actual data-generating process that gave rise to the data. It was emphasized that the nature of induction underlying the different forms of inference relating to estimation, testing and prediction is distinct; the reasoning underlying *estimation* and *prediction* is *factual* but that of testing is *counterfactual*.

The discussion brought out the importance of supplementing the pre-data error probabilities with a post-data severity assessment in order to bridge the gap between the coarse N-P test accept/reject decision and a data-specific inferential interpretation of the result. Supplementing N-P testing with the severity evaluation gives rise to what Mayo (1996) calls the *error-statistical account* of inference which constitutes the most well-rounded description of frequentist statistical induction.

An important aspect of statistical induction concerns the validity of the premises. The reliability of inference depends crucially on the validity of the statistical premises. When the statistical premises are misspecified the *actual* and *nominal* error probabilities (pre-data and post-data) differ, giving rise to unreliable inferences. The crucial role played by the notion of statistical adequacy in ensuring the statistical validity of the inference was discussed at some length, and several methodological issues pertaining to misspecification testing and respecification have been noted; severe testing reasoning plays a crucial role in understanding these issues. The paper also argues that the current literature on model selection ignores the statistical adequacy issue with dire consequences for the reliability of such procedures. Kepler's original data are used to illustrate the statistical vs. substantive adequacy issue by revisiting his first law of planetary motion.

References

[1] Berger, J. O. and T. Selke (1987) "Testing a Point Null Hypothesis: The Irreconciliability of P Values and Evidence (with discussion), *Journal of the American Statistical Association*, **82**: 112-122.

[2] Chatterjee, S. K. (2003) *Statistical Thought : A Perspective and History,* Oxford University Press, Oxford.

[3] Cox, D. R. (1990) "Role of Models in Statistical Analysis," *Statistical Science*, **5**: 169-174.

[4] Cox, D. R. and D. V. Hinkley (1974) *Theoretical Statistics*, Chapman & Hall, London.

[5] Cox, D. R. and N. Wermuth (1996) *Multivariate Dependencies: Models, Analysis and Interpretation*, CRC Press, London.

[6] Day, J. P. (1961) *Inductive Probability*, Routledge & Kegan Paul, London.

[7] Fisher, R. A. (1921) "On the 'Probable Error' of a Coefficient of Correlation Deduced from a small sample," *Metron*, 3-32.

[8] Fisher, R. A. (1922) "On the mathematical foundations of theoretical statistics", *Philosophical Transactions of the Royal Society A*, **222**: 309-368.

[9] Fisher, R. A. (1925) *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh.

[10] Fisher, R. A. (1935) *The Design of Experiments*, Oliver and Boyd, Edinburgh.

[11] Fisher, R. A. (1955) "Statistical methods and scientific induction," *Journal of the Royal Statistical Society*, B, 17: 69-78.

[12] Fisher, R. A. (1956) *Statistical methods and scientific inference*, Oliver and Boyd, Edinburgh.

[13] Glymour, C. (1981) *Theory and Evidence*, Princeton University Press, NJ.

[14] Hacking, I. (1965) *Logic of statistical Inference*, Cambridge University Press, Cambridge.

[15] Kieffer, J. (1977) "Conditional Confidence Statements and Confidence Estimators," *Journal of the American Statistical Association*, **72**: 789-827.

[16] Lawden, D. F. (2002) *Introduction to Tensor Calculus, Relativity and Cosmology*, Dover, New York.

[17] Lahiri, P. (2001) *Model Selection*, Institute of Mathematical Statistics, Ohio.

[18] Lehmann, E. L. (1986) *Testing statistical hypotheses*, 2nd edition, Wiley, New York.

[19] Lehmann, E. L. (1990) "Model specification: the views of Fisher and Neyman, and later developments", *Statistical Science*, **5**: 160-168.

[20] Lehmann, E. L. (1993) "The Fisher and Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *Journal of the American Statistical Association*, **88**: 1242-9.

[21] Mayo, D. G. (1991) "Novel Evidence and Severe Tests," *Philosophy of Science*, **58:** 523-552.

[22] Mayo, D. G. (1996) *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago.

[23] Mayo, D. G. and D. R. Cox (2006) "Frequentist statistics as a theory of inductive inference," pp. 96-123 in *The Second Erich L. Lehmann Symposium – Optimality*, Lecture Notes-Monograph Series, Volume 49, Institute of Mathematical Statistics.

[24] Mayo, D. G. and A. Spanos (2004) "Methodology in Practice: Statistical Misspecification Testing", *Philosophy of Science*, **71**: 1007-1025.

[25] Mayo, D. G. and A. Spanos (2006) "Severe Testing as a Basic Concept in the Neyman-Pearson Philosophy of Induction," forthcoming in *The British Journal for the Philosophy of Science*.

[26] Neyman, J. (1950) *First course in probability and statistics*, Henry Holt, New York.

[27] Neyman, J. (1952) *Lectures and conferences on mathematical statistics and probability*, 2nd ed. U.S. Department of Agriculture, Washington.

[28] Neyman, J. (1955) "The Problem of Inductive Inference," *Communications on Pure and Applied Mathematics*, **VIII**: 13-46.

[29] Neyman, J. (1956) "Note on an Article by Sir Ronald Fisher," Journal of the Royal Statistical Society. Series B (Methodological), **18**: 288-294.

[30] Neyman, J. (1957a) "Inductive Behavior as a Basic Concept of Philosophy of Science," *Revue Inst. Int. De Stat.*, **25**: 7-22.

[31] Neyman, J. (1957b) "The Use of the Concept of Power in Agricultural Experimentation," *Journal of the Indian Society of Agricultural Statistics*, **IX**: 9-17.

[32] Neyman, J. and E. S. Pearson (1933) "On the problem of the most efficient tests of statistical hypotheses", *Philosophical Transactions of the Royal Society*, A, 231: 289-337.

[33] Pearson, K. (1895) "Contributions to the mathematical theory of evolution II. Skew variation in homogeneous material", *Philosophical Transactions of the Royal Society of London, series A*, **186**, 343-414.

[34] Pearson, K. (1920) "The Fundamental Problem of Practical Statistics," *Biometrika*, **XIII**, 1-16.

[35] Pierce, C. S. (1931-5), Collected Papers, Vols. 1-6, ed. by Hartshorne and P. Weiss, Harvard University Press, Cambridge.

[36] Rao, C. R. (1992) "R. A. Fisher: The Founder of Modern Statistics," *Statistical Science*, **7**, 34-48.

[37] Rao, C. R. and Y. Wu (2001) "On Model Selection," pp. 1-64, in P. Lahiri (2001).

[38] Salmon, W. (1967) The Foundations of Scientific Inference. Pittsburgh, University of Pittsburgh Press.

[39] Spanos, A., (1986) *Statistical Foundations of Econometric Modelling*, Cambridge University Press, Cambridge.

[40] Spanos, A. (1995) "On theory testing in Econometrics: modeling with nonexperimental data", *Journal of Econometrics,* **67**:189-226.

[41] Spanos, A. (1999) *Probability Theory and Statistical Inference: econometric modeling with observational data*, Cambridge University Press, Cambridge.

[42] Spanos, A. (2005a) "Misspecification and the Reliability of Inference: the t-test in the presence of Markov dependence," Virginia Tech Working Paper.

[43] Spanos, A. (2005b) "Omitted Variables and Artificial Regressions: A General Approach to Misspecification Testing," Virginia Tech Working Paper.

[44] Spanos, A. (2005c) "Structural Equation Modeling, Causal Inference and Statistical Adequacy," pp. 639-661, *Logic, Methodology and Philosophy of Science: Proceedings of the Twelfth International Congress,* Editors, P. Hajek, L. Valdes-Villanueva and D. Westerstahl, King's College, London.

[45] Spanos, A. (2005d) "Structural vs. Statistical Models in Empirical Modeling: Kepler's first law of planetery motion revisited," Virginia Tech working paper.

[46] Spanos, A. (2006a) "Econometrics in Retrospect and Prospect," in the *Palgrave Handbook of Econometrics, vol. 1: Theoretical Econometrics,* London: MacMillan, pp. 3-58.

[47] Spanos, A. (2006b) "Where Do Statistical Models Come From? Revisiting the Problem of Specification," pp. 124-150 in *The Second Erich L. Lehmann Symposium – Optimality*, Lecture Notes-Monograph Series, Volume 49, Institute of Mathematical Statistics.

[48] Spanos, A. (2006c) "The Curve-Fitting Problem, Akaike-type Model Selection, and the Error Statistical Approach." Virginia Tech working paper.

[49] Spanos, A. and A. McGuirk (2001) "The Model Specification Problem from a Probabilistic Reduction Perspective," *Journal of the American Agricultural Association*, **83**, 1168-1176.

[50] Worrall, J. (2002) "New Evidence for Old' in P.Gardenførs et al (eds), *The Scope of Logic, Methodology and Philosophy of Science*, Kluwer, Dordrecht.